Parallel Computing 000 (2017) 1–12

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

Energy balance between voltage-frequency scaling and resilience for linear algebra routines on low-power multicore architectures

Sandra Catalán[a], José R. Herrero[b,*], Enrique S. Quintana-Ortí[a], Rafael Rodríguez-Sánchez[a]

[a] Depto. Ingeniería y Ciencia de Computadores, Universidad Jaume I, Castellón, Spain
[b] Dept. d'Arquitectura de Computadors, Universitat Politècnica de Catalunya, Spain

ARTICLE INFO

Article history:
Received 15 July 2016
Revised 25 April 2017
Accepted 20 May 2017
Available online xxx

Keywords:
Energy efficiency
Voltage-frequency scaling
Fault tolerance
Dense linear algebra
High performance
Multicore processors

ABSTRACT

Near Threshold Voltage (NTV) computing has been recently proposed as a technique to save energy, at the cost of incurring higher error rates including, among others, Silent Data Corruption (SDC). In this paper, we evaluate the energy efficiency of dense linear algebra routines using several low-power multicore processors and we analyze whether the potential energy reduction achieved when scaling the processor to operate at a low voltage compensates the cost of integrating a fault tolerance mechanism that tackles SDC. Our study targets algorithmic-based fault-tolerant versions of the dense matrix-vector and matrix(-matrix) multiplication kernels (GEMV and GEMM, respectively), using the BLIS framework, as well as an implementation of the LU factorization with partial pivoting built on top of GEMM. Furthermore, we tailor the study for a number of representative 32-bit and 64-bit multicore processors from ARM that were specifically designed for energy efficiency.

1. Introduction

Aggressive technology scaling is steadily shrinking transistor size, at the pace dictated by Moore's Law [1], increasing the occurrence of faults in computing systems in the way. In the future, as the number of components integrated in CMOS circuits grows, the mean time between failures (MTBF) for the full system will be significantly reduced, promoting *resilience* into a crucial challenge on the road towards Exascale systems [2,3].

Under the pressure of the *energy wall* [2–4], near threshold voltage (NTV) computing has been proposed as a means to reduce energy consumption [5]. However, scaling the operational voltage diminishes the critical charge required to flip a stored value [6,7], so that particles of low energy due to atmospheric radiation can cause soft errors (SE). Consequently, error rates for low power modes are higher than those present in high power modes, as the error rate grows exponentially with the reduction on the supply voltage [8].

Already today, reliability and energy consumption are key criteria for system design. However, hardware protection mechanisms cannot be relied upon as the sole method for SE mitigation. Instead, they should be used in conjunction with other

* Corresponding author.
E-mail addresses: catalans@uji.es (S. Catalán), josepr@ac.upc.edu (J.R. Herrero), quintana@uji.es (E.S. Quintana-Ortí), rarodrig@uji.es (R. Rodríguez-Sánchez).

http://dx.doi.org/10.1016/j.parco.2017.05.004
0167-8191/

mitigation techniques for improved reliability at an energy-efficient cost. In this scenario, fault tolerance techniques such as checkpointing or redundancy (replication) may be too costly from the computational and energy perspectives, favoring alternative approaches based on application-specific algorithmic-based fault tolerance (ABFT) [3,9,10].

In this paper, we investigate the energy costs of tackling in software with SE that result in silent data corruption (SDC), using the dense matrix-vector multiplication (GEMV) and dense matrix multiplication (GEMM) as case studies. These two kernels are the cornerstone upon which the entire dense linear algebra (DLA) software stack, and implicitly many scientific and engineering codes, rely for high performance. One particular example is the LU factorization with partial pivoting (GETRF), which can be encoded to cast a major fraction of its computations in terms of GEMM, and is adopted in our work as an additional case study. As target architectures, we employ several general-purpose multicore processors from ARM that are specially designed to deliver reasonable performance with high energy efficiency. In more detail, our paper makes the following contributions:

- We evaluate the energy efficiency under different voltage-frequency scaling (VFS) configurations. For this purpose, we leverage efficient multi-threaded implementations of GEMV, GEMM and GETRF based on the BLIS framework [11], especially tailored for the ARMv7 Cortex-A7/A15 and the ARMv8 Cortex-A53/A57.
- We review the theoretical costs of introducing simple software resilience techniques for GEMV, GEMM, and GETRF. For GEMV we discuss how to deal with the errors via redundancy, exploiting the memory-bound nature of this kernel to deliver an affordable resilience mechanism. For GEMM, we follow the ABFT described in [12]; and the same mechanism provides a reliable solution for GETRF, when built on top of a resilient GEMM.
- Finally, we introduce isoenergy models that capture the interplay between VFS, error detection costs, and error correction overhead/error rates, and how these factors combine to modify the energy efficiency for these three linear algebra operations and the target low-power architectures when operating in *low-voltage at extended margins* (LVEMs) configurations.

At this point, we note that the detection+correction strategies discussed for the target dense linear algebra operations can be regarded as theoretical proposals, and the details of their actual implementation do not impact our isoenergy model. Indeed, we believe that a detailed analysis of the reliability of the error detection (and correction) mechanism(s), which may indeed suffer from errors themselves, is beyond the scope for this work. We consider this as a clear target for a research that aims to produce practical and efficient resilient implementations of dense linear algebra kernels; see [12].

In Section 2 we review some related work. In Section 3 we briefly review the BLIS implementation of GEMV and GEMM, and we describe how to modify them in order to produce resilient versions, as well as the implications on the LU factorization in LAPACK. Section 4 outlines the experimental setup. In Section 5 we present an experimental evaluation of four ARM multicore processors from the points of view of performance and energy efficiency. Next, in Section 6 we analyze the trade-off between energy consumption and fault tolerance. Finally, in Section 7 we summarize our work with some concluding remarks.

2. Related work

Reducing the operating voltage of an electronic circuit is a well-known technique that can potentially diminish its power consumption. Oftentimes, this reduction in voltage comes together with a decrease of the operational frequency, an strategy known as VFS [13–15]. However, the supply voltage can also be reduced while maintaining the operating frequency, an approach known as undervolting, in an attempt to save power while preserving throughput. Undervolting and VFS outside of the nominal region can nevertheless introduce errors, which need to be corrected in case the voltage is dropped in excess. Hardware support for error detection/correction from operation at lower supply voltage was introduced in [16]. The impact on the memory system was specifically studied in [17–19]. In [20] the authors explore the potential energy benefits of reducing the chip's voltage to the safe limit ($V_{\min}$) at a fixed frequency. ($V_{\min}$ is program dependent.) Exceeding such safe limits causes SDC to arise, with an avalanche error effect when the voltage is pushed below a certain threshold. Interestingly, an additional 4–5% undervolt below $V_{\min}$ causes the OS to crash. However, that work does not address resilience mechanisms. Instead, it focuses only on shifting the guard-band down for energy improvement without impacting the correctness level. They show that there is about a 20% voltage guard-band for the graphics processors tested in their work which can result in up to 25% energy savings.

While most previous research focuses on exploring energy-saving and resilience-enhancing opportunities separately, only very few selected publications study their interactions. Some works assess the energy costs of traditional resilience-enhancing methods such as checkpointing/restart or replication. The former is analyzed in [21], modeling its energy costs and introducing checkpoint compression to reduce the energy consumption. The latter is studied in [22], which examines the energy costs of coordinated checkpointing and replication, contributing mainly to make replication more time- and energy-efficient. In [10,23] the authors study the interplay between energy efficiency and resilience (mainly the checkpoint/restart technique) in high performance computing with a focus on undervolting. More specifically, they develop analytical models to investigate the potential of achieving high energy efficiency in HPC by undervolting, with hardware/software-level resilience techniques applied on-the-fly to guarantee the correct execution of HPC runs.

In this work we consider low-voltage at extended margins (LVEM), i.e., configurations operating below nominal voltage which may incur soft errors, but always considering voltage values above some threshold where an avalanche error effect

Please cite this article as: S. Catalán et al., Energy balance between voltage-frequency scaling and resilience for linear algebra routines on low-power multicore architectures, Parallel Computing (2017), http://dx.doi.org/10.1016/j.parco.2017.05.004